Convolutional Neural Networks for Steel Surface Defect Detection from Photometric Stereo Images

D. Soukup and R. Huber-Mörk

Safety and Security Department,
AIT Austrian Institute of Technology GmbH, Austria

Abstract. Convolutional neural networks (CNNs) achieved impressive recognition rates in image classification tasks recently. In order to exploit those capabilities, we trained CNNs on a database of photometric stereo images of metal surface defects, i.e. rail defects. Those defects are cavities in the rail surface and are indication for further surface degradation right up to rail break. Due to security issues, defects have to be recognized early in order to take countermeasures in time. By means of differently colored light-sources illuminating the rail surfaces from different and constant directions, those cavities are made visible in a photometric dark-field setup. So far, a model-based approach has been used for image classification, which expressed the expected reflection properties of surface defects in contrast to non-defects. In this work, we experimented with classical CNNs trained in pure supervised manner and also explored the impact of regularization methods such as unsupervised layer-wise pre-training and training data-set augmentation. The classical CNN already distinctly outperforms the model-based approach. Moreover, regularization methods yet yield further improvements.

1 Introduction

Machine vision for quality control is an established field in industrial inspection, which is characterized by rather well defined acquisition conditions and elaborated expert knowledge about the universe of possible defects to be detected. Mobile machine vision, in our case the detection of rail surface defects from a mobile platform, i.e. a rail car, becomes more difficult w.r.t. imaging conditions and variation of defect appearance. On the other hand, over the last decade, computer vision made impressive progress in the field of more general detection and classification problems such as face detection, object classification, text recognition, just to name a few application areas. Very recently, the *renaissance* of neural network based methods delivered impressive results on benchmark and real world data sets for digit recognition [1], traffic sign recognition [2] and privacy protection by face and number plate detection from Google StreetView images [3].

Speaking about typical defect detection scenarios in machine vision, the comparison to one or a set of so called *master* image(s), requires to have access to a well defined model of non-defective parts or objects. Furthermore a reasonable

G. Bebis et al. (Eds.): ISVC 2014, Part I, LNCS 8887, pp. 668–677, 2014.

metric is necessary which defines whether a deviation from the master image(s) is acceptable or not. For goods with well defined geometric, i.e. machine parts, or color properties, e.g. print products, tolerances could be defined by requirements. For applications such as the considered one, i.e. surface defect detection on rails from a mobile platform, the knowledge of origin and appearance of the defects is not fully understood. There exists the theory that rail surface defects, which are essentially small holes called *headchecks* or *spallings*, result from rail wear induced by *rolling contact fatigue* (RCF) [4]. Those cavities are assumed to emerge from crossing sections of small microcracks on the rails' surface. Modeling the image appearance of those defects is based on simplifications and assumptions. Defect modeling and detection using spatial correlation statistics [5] and co-occurrence based descriptors [6] was applied to the task at hand. An approach based on learning defect detection from data should at first better represent the data and, finally, increase detection rates.

The appearance of metal surfaces in machine vision depends both on material properties as well as illumination and acquisition geometry. Appropriate relative placement of illumination sources, cameras and observed objects enhances specific object properties, e.g. texture, edges, surface deformations etc., and assists automatic analysis methods. Active illumination offers a variety of possibilities with respect to geometrical arrangement, spectral properties and structuring of light (e.g. pattern projection). Geometrical arrangement of light sources enables 2.5D to 3D analysis approaches such as shape from shading [7] and photometric stereo [8], where images under different illumination are used to reconstruct the surface under investigation or, at least, to infer surface characteristics.

After a discussion of related work in Section 2, we describe the approaches, the model-based as well as the CNNs, in Section 3. Classification results are presented in Section 4, followed by the conclusion in Section 5.

2 Related Work

Recent progress in the field of neural networks was triggered by a number of achievements. *Deep learning architectures* were enabled by increasing computing power and provide more and higher levels of representation [9]. Data augmentation, e.g. addition of artificial training data derived from the existing data through distortions, proved to be a powerful tool to avoid overfitting [10]. Committee methods are able to reduce the error rate by combination of several networks, especially when the individual predictions are uncorrelated [1]. Finally, unsupervised methods for learning features and representations became very popular and solved problems with purely supervised training, e.g. dependence on random initialization, slow convergence etc. [11].

The convolutional neural network (CNN) is a neural network architecture especially designed for image recognition [12]. Outstanding performance of CNNs was reported for benchmark problems including the MNIST handwritten digit recognition [1], the Google StreeView house number (SVHN) data set [3] and the German traffic sign recognition benchmark [2]. CNNs are multilayer neural

network architectures implementing local receptive fields through *convolutional layers* and invariance w.r.t. small geometric deformations through *pooling layers*. CNNs are usually trained in supervised fashion, although the Neocognitron [13], which could be regarded as a predecessor of the CNN, was suggested to be trained in self-organized fashion. Recently, the application of CNNs for classification of seven different types of steel defects was described [14].

For metal surfaces the relationship between specular reflection and diffuse scattering depends primarily on the light wavelength and surface roughness [15]. For precisely polished metal surfaces we could expect that the amount of reflected light predominantly depends on the placement of light source and observer given by the law of reflection (incoming angle equals outgoing angle with respect to surface normal). For grinded metal surfaces some diffuse scattering component will be overlaid. Concerning the illumination setup, in bright-field illumination the goal is to direct the illumination from the observed object via reflection to the sensor, while reflection from the object towards the sensor is avoided in a dark-field setup. The typical application of a dark-field setup is to direct reflection towards an image senor in order to detect anomalities, e.g. discontinuities, edges or distortions. Anomaly detection in milled steel surfaces using various approaches for 2D texture analysis was discussed [16]. A method using a series of bright- and dark-field illuminated images and 2.5D analysis was used for solder paste inspection [17]. An experimental study describes specular reflection from metallic surfaces and application to coin classification using images illuminated by light sources differing in location and spectral emission [18].

3 Approach

In this section we discuss the arrangement of lighting and camera components to obtain surface disruptions from a photometric stereo arrangement. Due to the common lack of benchmarks for very specialized tasks such as the one discussed in this paper, we compare the CNN based approach to a model based approach motivated by photometric considerations of reflection.

3.1 Photometric Acquisition Setup

A dark-field acquisition setup is used to make deviations from a continuous surface visible, i.e. illumination is only scattered back to the sensor in regions forming cavities or small hills, see Fig. 1(a). In order to discriminate between cavities and hills one makes use of spatially localized reflection behavior around the region in question. Fig. 1(b) shows an idealized cavity to be detected by analyzing the corresponding sensor's responses from the edges of the cavity as shown in Fig. 1(c).

Images obtained with the setup from Fig. 1 are shown in Fig. 2. Various examples for cavities are shown in Fig. 2(a), whereas Fig. 2(b) shows examples from other areas such as non-defective or grinded surface as well as regions showing microcracks.

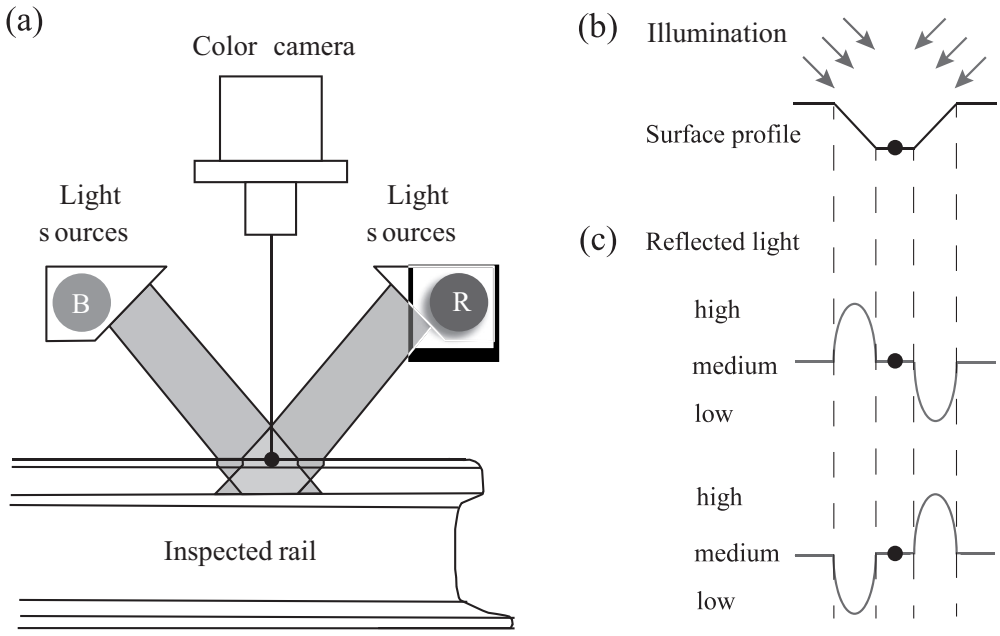

Fig. 1. Acquisition setup and model of reflection properties: (a) top-down view of the head surface using a line camera and illumination by different line light sources under oblique angles, (b) distorted surface profile and direction of illumination sources, (c) model of reflectance for a distorted surface profile

3.2 Model Based Detection

Referring to Fig. 1(c) we define the *detection position* to be in the middle of the surface profile, i.e. the black dot in Figs. 1(a)-(c). We denote the red and blue light received by the camera r and b and define four heuristic target detection equations

$$r_{+d}/\bar{r} < 1/t_r \ , \quad r_{-d}/\bar{r} > t_r \ , \quad b_{-d}/\bar{b} < 1/t_b \ , \quad b_{+d}/\bar{b} > t_b \ . \tag{1}$$

The subscript $-d$ indicates a positional offset in rail direction, e.g. at a distance of d pixels to the left, with respect to the detection position. The position $+d$ is located d pixels to the right w.r.t. the detection position. The r_i, b_i are the red and blue values for the pixel at offset position $i \in \{-d, +d\}$ and $\bar{r}, \bar{b}$ are background estimated by local smoothing. Two different thresholds t_r, t_b, related to the different color channels, were derived empirically in order to achieve optimal detection rates with little false positives. The combination of the four target detectors is done by taking into account the anti-correlation property, i.e. the presence of a target is assumed only if at least three detectors coincide in their decision for a target.

3.3 Learned CNN Detector

A CNN is a multi-layer neural network architecture that is especially suited for the processing of image data. There are three kinds of layers, convolutional

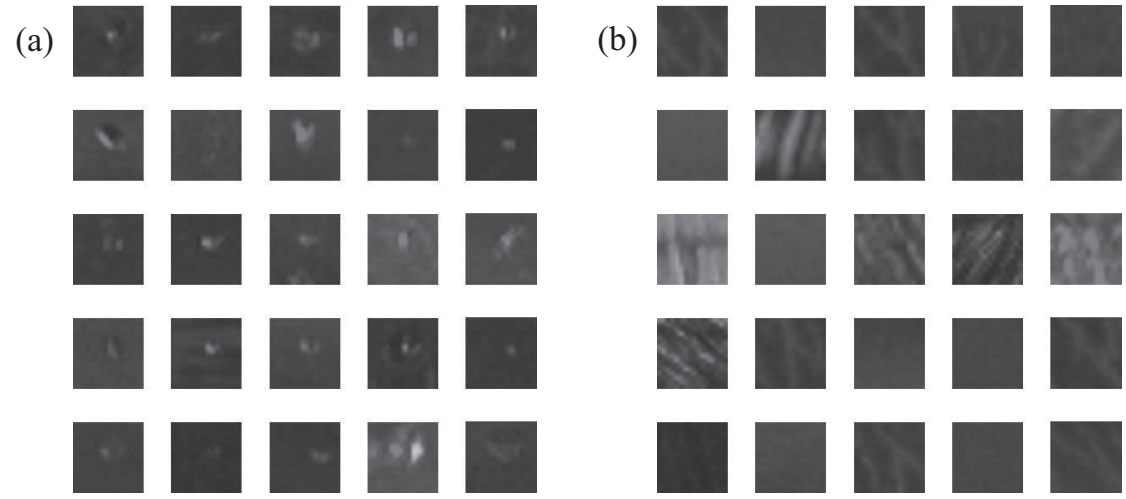

Fig. 2. Image patches acquired by a dark-field setup and photometric illumination: a) examples for surface defects, b) non-defective samples

layers, pooling layers, and at least one final fully connected layer. A convolutional layer consists of a filter bank and produces a fixed number of output images (output maps) from a number of input images (input maps). The input maps are filtered with the corresponding filters, all filter results related to the same output map are aggregated (e.g. summed) and a learnable bias value is added. Finally, the output maps are generated by applying a non-linear function (e.g. tanh) to all output map pixels. The filter coefficients can be viewed as weights of neurons, which are represented by the pixels of the output maps. In the training process, these weights together with the bias values are learned by means of back propagation. Due to the convolutional processing, all neurons (i.e. pixels) of an output map have the same weights in form of the corresponding filter values. This property is called weight sharing and makes CNN training efficient. Each convolutional layer is followed by a so called pooling layer, in which pixels within output maps (neurons) are locally aggregated (e.g. max filtering) and the maps are downsampled. While the convolutional processing captures the spatial relation between neurons, which represent image features, the pooling step introduces more spatial tolerance w.r.t. feature positions in the images. Additionally, the downsampling step in the pooling layers successively engenders larger receptive fields in following convolutional layers. While the neurons in the first convolutional layer are sensitive to very local image structures like edges, neurons in deeper convolutional layers more and more learn combinations of lower level features. Eventually, the last pooling layer is followed by one fully connected layer with a number of output neurons corresponding to the number of image classes that have to be recognized.

We trained CNNs with two convolutional layers with 6 and 12 output maps, respectively (see Fig. 3). All filter sizes were chosen to be 5×5. The pooling layers subsequently to the convolutional layers both accomplish 2×2 max-pooling and downsampling by a factor of 2. The final fully connnected layer has two

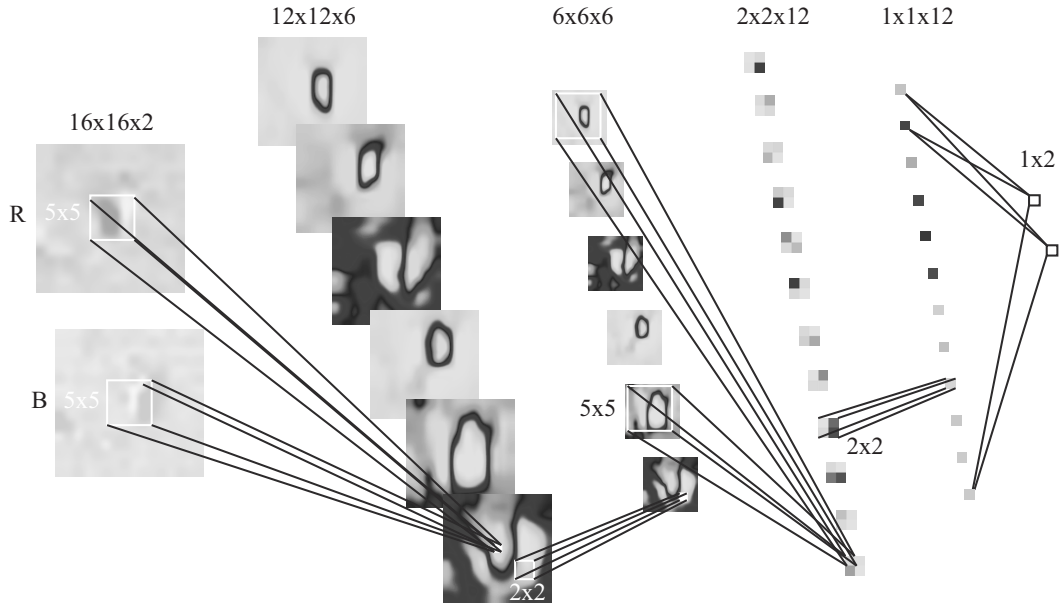

Fig. 3. CNN architecture for surface defect detection: two convolutional and pooling layers and a final fully connected layer

output neurons, one indicating surface defects, the other non-surface defects. Fig. 3 shows the used CNN architecture along with images of input and layer activations.

Usually, CNNs are trained in supervised manner. However, unsupervised layer-wise pre-training has also been suggested (e.g. [19]) as a regularization method. In order to investigate the influence of CNN pre-training on the error rates, we initialized the convolutional layers' filter banks with the weights of sparse auto-encoders trained in unsupervised manner on all 5×5 image patches of corresponding input maps.

Moreover, we used data augmentation of training data sets as another regularizer, where the sizes of the sample sets were increased by a certain factor. For this purpose, a number of random pixel positions in each sample was displaced by a random, but bounded offset and the corresponding distorted sample image was obtained by warping using displacement vector fields interpolated by thin plate splines [20].

4 Results

We evaluated the CNN based approach by comparing it to a model based approach with handcrafted parameter adjustment. A total number of 2532 image positions showing cavities and non-cavities were manually identified. Patches of size 16×16 pixels were extracted from color images with a pixel resolution of approximately 0.23mm.

Five-fold cross-validation was used for the estimation of detection rates in the learned approach. The model-based approach is assessed by ROC-analysis.

4.1 Model Based Detector

With the model-based approach one can adjust the two threshold parameters t_r and t_b from Eq. 1 to obtain varying rates of true and false positives. Fig.4(a) shows two ROC curves, one for t_r fixed at the value were the sum of false positive and negative decisions is at a minimum and one curve for t_b fixed under the same condition, respectively. For the chosen point of operation an error rate of 7.11% was achieved. A number of false negatives is shown in Fig.4(b), and Fig.4(c) shows some examples for false positives.

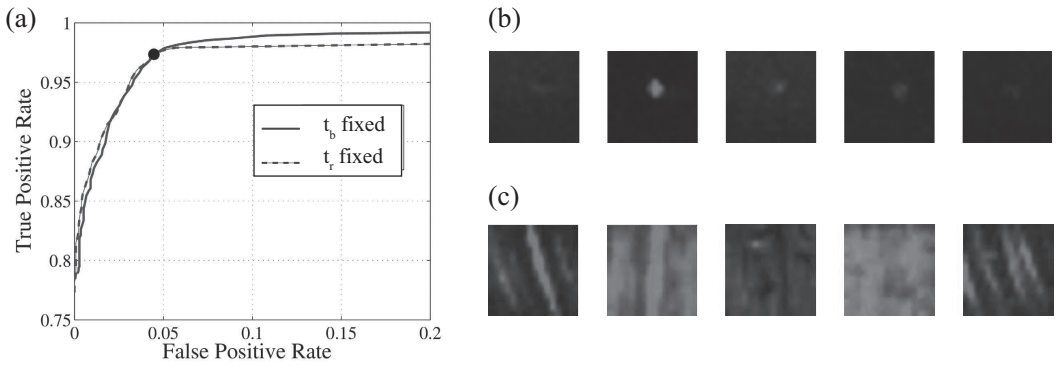

Fig. 4. Model based detection results: (a) ROC curves for varying thresholds in blue and red channels, examples for (b) false negatives and (c) false positives

4.2 CNN Based Detector

Each CNN was trained in supervised manner for 150 epochs with a constant learning rate of $\alpha = 10$ and mini-batch size of 75. To determine the influence of unsupervised pre-training on the error rates, we trained randomly initialized CNNs in pure supervised manner as well as CNNs, where the convolutional layers' filters were initialized with weights from sparse auto-encoders trained in unsupervised manner on corresponding 5×5 patches of the layers' input maps. The auto-encoders' weight-decay parameter was set to $\lambda = 0.00001$, the sparsity parameter $\rho = 0.5$, and $\beta = 0.5$ (influence of sparsity term in optimization objective function). The maximal number of iterations of the auto-encoder optimization procedure was set to 5000, however, they all converged earlier.

Each CNN was trained 3 times on each fold of the five-fold data set and the minimum error rate CNN of these 3 runs was used for cross validation. Eventually, the five-fold cross-validation error rates were 1.108% for randomly initialized CNNs and 0.67% for auto-encoder pre-trained CNNs. This result shows clearly that unsupervised, layer-wise pre-training had a significant positive impact on the recognition performance of the CNNs.

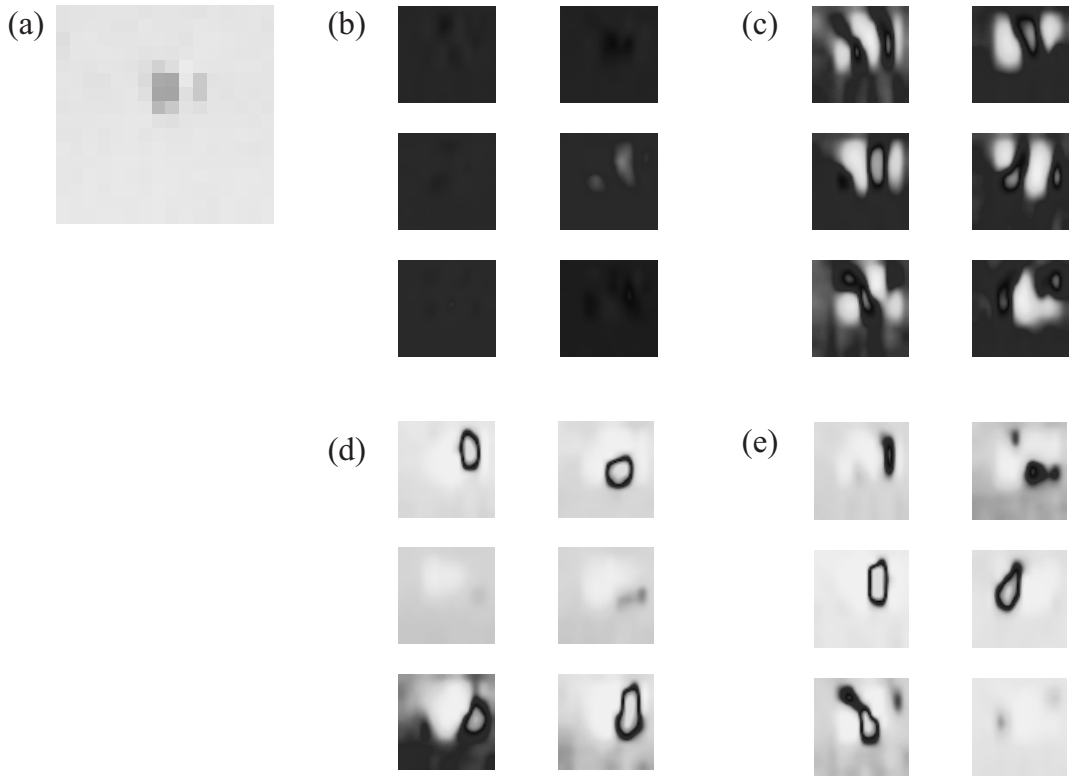

Fig. 5. Sample defect and corresponding six output maps of the first convolutional layer. (a) sample defect, (b) random initialization, (c) unsupervised pre-trained initialization (auto-encoders), (d) randomly initialized maps after 150 epochs, and (e) auto-encoder initialized maps after 150 epochs. Note, that stuctures in b) are still contained in d), while structures in a) are not perceivable in (d).

By means of the output maps of the first convolutional layers of a randomly initialized and an unsupervised pre-trained CNN (see Fig. 5), respectively, it can be perceived, that the basic pre-trained filter structures survive those 150 epochs of supervised CNN training. Apparently, the supervised training task of the CNNs is set to a meaningful region in parameter space by the auto-encoders' weights, such that the optimization procedure at least does not get caught in some random local sub-optimal minimum. Anyhow, this does not mean, that a single randomly initialized CNN could never be better. Occasionally, random initialization yields a better CNN, but not in the long term when trained repetitively.

A second CNN experiment was run, similar to the first one, but this time the corresponding folds serving as training data were increased by a factor of 3 by means of data augmentation. In these cases, the cross-validation error rates were 0.558% for the randomly initialized CNNs and 0.556% for the unsupervised pre-trained CNNs. We see that data augmentation also significantly improves the recognition performance of both kinds of CNNs, even more than unsupervised pre-training. However, unsupervised pre-training obviously could not improve the performance of CNNs trained on augmented training data any further.

Our data set is quite small, thus the training is prone to overfitting. It seems that unsupervised, layer-wise pre-training sets the starting point for supervised CNN training to a meaningful region in parameter space, thus avoiding optima due to overfitting. On the other hand, data augmentation prevents overfitting by increasing variation in the training data. Apparently, both methods improve the recognition performance by overcoming the drawbacks of small training data sets.

5 Conclusion

We trained CNNs on photometric stereo images of rail surfaces in a dark-field setup in order to detect rail surface defects. Up to now this classification task has been accomplished by means of a model-based approach. However, the detection performance of the CNNs showed to be significantly better than the model-based approach's. Moreover, we could demonstrate how overfitting due to our relatively small training data set could be alleviated by the use of regularization methods, i.e. unsupervised layer-wise pre-training and training data augmentation. Regularization further improved the recognition rates. In future work, we want to extend the classification task to more classes, in order to be able to recognize different types of defective and non-defective rail surface structures, whereby deeper CNNs may be required.

References

1. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Proc. of Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3642–3649 (June 2012)
2. Ciresan, D., Meier, U., Masci, J., Schmidhuber, J.: A committee of neural networks for traffic sign classification. In: Proc. of International Joint Conference on Neural Networks (IJCNN), pp. 1918–1921 (July 2011)
3. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. In: Proc. of International Conference on Learning Representations (ICLR) April
4. Doherty, A., Clark, S., Care, R., Dembowsky, M.: Why rails crack. Ingenia (23), 23–28 (2005)
5. Huber-Mörk, R., Nölle, M., Oberhauser, A., Fischmeister, E.: Statistical rail surface classification based on 2D and $2^1/_2$D image analysis. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010, Part I. LNCS, vol. 6474, pp. 50–61. Springer, Heidelberg (2010)
6. Soukup, D., Huber-Mörk, R.: Cross-channel co-occurrence matrices for robust characterization of surface disruptions in $2^1/_2$D rail image analysis. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P., Zemčík, P. (eds.) ACIVS 2012. LNCS, vol. 7517, pp. 167–177. Springer, Heidelberg (2012)
7. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical Engineering 19(1), 139–144 (1980)
8. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. International Journal of Computer Vision 72(3), 239–257 (2007)

9. Arel, I., Rose, D.C., Karnowski, T.P.: Deep machine learning - a new frontier in artificial intelligence research. IEEE Computational Intelligence Magazine 5(4), 13–18 (2010)
10. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: Proc. of International Conference on Document Analysis and Recognition (ICDAR), pp. 958–963 (2003)
11. Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. In: Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS) (2011)
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
13. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36(4), 193–202 (1980)
14. Masci, J., Meier, U., Ciresan, D., Schmidhuber, J., Fricout, G.: Steel defect classification with max-pooling convolutional neural networks. In: Proc. of International Joint Conference on Neural Networks (IJCNN), pp. 1–6 (June 2012)
15. Westin, S.H., Li, H., Torrance, K.E.: A comparison of four BRDF models. In: Jensen, H.W., Keller, A. (eds.) Proc. of Eurographics Symposium on Rendering, pp. 1–10 (2004)
16. Herwig, J., Leßmann, S., Bürger, F., Pauli, J.: Adaptive anomaly detection within near-regular milling textures. In: Proc. International Symposium on Image and Signal Processing and Analysis, Trieste, Italy, pp. 106–111 (2013)
17. Pang, G.K.H., Chu, M.-H.: Automated optical inspection of solder paste based on 2.5D visual images. In: Proc. of International Conference on Mechatronics and Automation, pp. 982–987 (2009)
18. Hoßfeld, M., Chu, W., Adameck, M., Eich, M.: Fast 3D-vision system to classify metallic coins by their embossed topography. Electronic Letters on Computer Vision and Image Analysis 5(4), 47–63 (2006)
19. Ciresan, D.C., Masci, J., Meier, U., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proc. of International Conference on Artificial Neural Networks, ICANN (2011)
20. Bookstein, F.L.: Principal warps: Thin plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. 11(6), 567–585 (1989)